MEASUREMENT, *4*(3), 135–155

FOCUS ARTICLE

Eight Decades of Measurement in Depression

Darcy A. Santor
School of Psychology
University of Ottawa, Ontario, Canada
Provincial Centre of Excellence in Child and Youth Medical Health
Ottawa, Ontario, Canada

Michelle Gregus
Department of Psychology
University of Alberta, Alberta, Canada

Andrew Welch
Department of Psychology
Dalhousie University, Halifax, Nova Scotia, Canada

Since 1918, more than 280 measures of depressive severity have been developed and published. These measures differ in content, response format, and objectives. This article examines (a) the characteristics of scales developed in the past 80 years, and (b) the frequency with which different scales have actually been used in basic science and treatment outcome studies over a 10-year period of time. Results of the authors' item analysis showed considerable variability across measures, but few consistent differences in item content emerged over time, other than in how anxiousness, suicidality, and loss of interest were assessed. Less than half of published scales assessed social withdrawal specifically, and as many as 20% of measures did not assess either "depression" or "sadness" directly. Worthlessness was assessed as thoroughly as depressed mood and more thoroughly than all other core symptoms of depression.

Correspondence should be addressed to Darcy A. Santor, School of Psychology, University of Ottawa, Lamoureaux Hall, 145 Jean-Jacques-Lussier St., P. O. Box 450, Station A, Ottawa, Ontario, K1N 6N5, Canada. E-mail: dsantor@uottawa.ca

Results of frequency analysis showed that despite the large number of scales developed to date, relatively few scales are actually used. Far more measures are used to assess depressive severity in basic science studies than in treatment studies of depression. Treatment studies have relied primarily on 6 measures of depression, the majority of which were developed 20 years ago and which are not representative of the larger body of recently developed measures of depressive severity measures. Implications for progress in the assessment of depression are discussed.

Keywords: depression, scales, measurement

In 1918, Woodworth developed what is arguably the first published measure of depression (Garret & Schenk, 1928). The Psychoneurotic Inventory operationalized the construct of neurasthenia, which consisted of a broad collection of depressive and anxiety symptoms. This measure was designed to identify American army recruits who were likely to experience "difficulties in adjusting themselves to the exigencies of military life" (Garret & Schenk, 1928, p. 459). Since then, the number of scales, inventories, and checklists developed to assess the severity of depression has increased substantially, with new measures of depression appearing each year (see Figure 1).

There are good reasons for constructing different measures of depression. Depression has been conceptualized in different ways, including psychodynamic

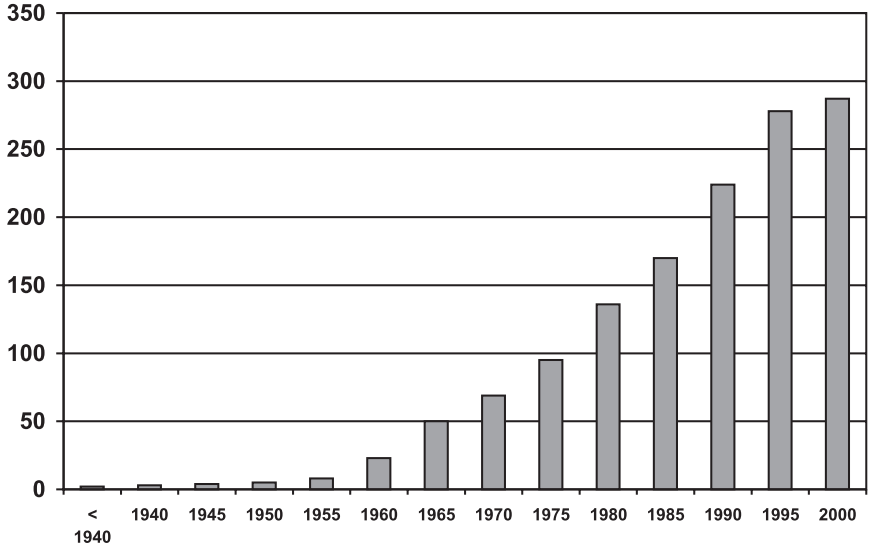

FIGURE 1 Cumulative number of measures of depression created in the past 80 years. Results show a steady rate of growth since 1960. Only in the past 5 years has the rate at which new measures of depression are created begun to slow.

(Arieti & Bemporad, 1980; Blatt, 1974; Freud, 1917), cognitive (Beck, 1963; Bibring, 1953), interpersonal (Coyne, 1976), behavioral (Lewinsohn, Youngren, & Grosscup, 1979), attachment (Bowlby, 1969), and evolutionary frameworks (Gilbert, 1992; Price, Sloman, Gardner, Gilbert, & Rohde, 1994), and may be expressed differently in different populations and contexts. Indeed, the abundance of measures of depression has made selecting appropriate measures difficult and evaluating differences among measures complex.

The numerous psychometric studies, reviews, and compendia of measures that have appeared in the past several years have sought to assist researchers and clinicians (cf. Nezu, McClure, Meadows, & Ronan, 2000) in selecting and evaluating measures of depression, either by comparing measures on the basis of various performance indexes (e.g., Beck, Steer, & Garbin, 1988) or on the basis of their differences in item content (e.g., Snaith, 1993). Snaith, for example, highlighted the considerable heterogeneity in item content in some of the most frequently used measures of depression, including the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960), the Beck Depression Inventory (BDI; Beck, Steer, & Brown, 1996; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), and the Center for Epidemiologic Studies Depression Scales (CES-D; Radloff, 1977). Reviews of this kind are important in helping researchers select measures prudently. Results of this review showed that the BDI, for example, contains a relatively higher proportion of cognitive items, whereas the HRSD contains a relatively greater proportion of somatic items (Snaith, 1993). Unfortunately, measures that are not studied, reviewed, or listed in compendia explicitly are unlikely to be considered widely. One objective of this study was to build on the review of Snaith and examine the extent to which item content varies across measures of depression in general. A second objective was to examine the manner in which measures of depression are most frequently used by directly estimating the frequency of use of depression measures over a 10-year period of time.

ITEM CONTENT

Item content can differ in a variety of ways, with respect to (a) the types of symptoms assessed, (b) the number of items used to assess an individual symptom, and (c) the proportion of items allocated to assessing a single symptom domain. Understanding the manner in which depressive severity has been operationalized across different scales provides one measure of the degree to which explicit and implicit concepts of depression differ across scales or have changed over time.

Historically, different positions have been articulated concerning the degree to which correspondence between definitions of constructs and operational criteria should and can be achieved. According to Bridgeman (1927), what is meant by a concept or construct "is synonymous with the corresponding set of operations" (p.

5). The strong view of this requirement, which Bridgeman intended, implies that different experimental procedures (or operations) correspond to different concepts or constructs. However, this view has been criticized as practically and theoretically problematic in its strong form. Indeed, some degree of equivalence must be allowed if science is to proceed (Carnap, 1936, 1937; Suppes, 1977). However, the other extreme—treating measures as equivalent—is equally problematic, which has been the focus of considerable debate (Coyne, 1994; Flett, Vredenburg, & Krames, 1997; Kendall & Flannery-Schroeder, 1995; Kendall, Hollon, Beck, & Ingram, 1987; Santor & Coyne, 2001a, 2001b; Tennen, Hall, & Affleck, 1995a, 1995b; Vredenburg, Flett, & Krames, 1993).

The review by Snaith (1993) showed that item content was extremely heterogeneous among some of the most commonly used measures of depressive severity. Accordingly, it is fair to expect that the types of symptoms assessed would vary across symptom domains for measures in general and not just a select few (H1). Indeed, there are a number of factors influencing how depressive severity is operationalized, which may account for the considerable variability among measures. These include the idiosyncratic interests of individual researchers, the specific conceptual framework in which depression is understood, and the target group of interest. However, differences in item content among scales are not necessarily problematic and in some instances should be anticipated. For example, irritability is considered diagnostic of depression in children but it is not included among those symptoms that are characteristic of depression in adults, which implies specificity in content across adult and pediatric measures of depressive severity. Accordingly, it can be hypothesized that a greater proportion of adolescent measures of depression should assess irritability than adult measures of depression (H2).

Formal definitions, diagnostic criteria, and the degree to which certain clusters of symptoms are emphasized, such as worthlessness, have changed somewhat over the past 80 years. Descriptions of depression in the first and second versions of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*; American Psychiatric Association [APA] 1952, 1968) emphasized anxious features that inventories such as the one developed by Woodworth in 1918 tried to capture. Later versions of the *DSM* (APA, 1980, 1987, 1994) explicitly separated depressive and anxious features. Accordingly, one might anticipate that the prominence of anxious symptoms on measures of depression has decreased over time (H3).

In addition, theorists, such as Bibring (1953) and Beck (1963), have emphasized the cognitive features of depression, such as a negative thinking, dysfunctional attitudes, and deficits in self-worth. Given the impact of cognitive theories of depression on the field, one might hypothesize that cognitive symptoms such as worthlessness may now be as frequently or thoroughly assessed as other core features of depression, such as mood or anhedonia (H4), and may have become more thoroughly assessed over time (H5).

FREQUENCY OF USE

There are also good reasons for examining the frequency with which measures of depression are actually used in practice. Indeed, what we know about depression depends on the quality of the measures used to assess depression. Many of the scales, such as the BDI, HRSD and CES-D—identified by Snaith (1993) as the most frequently used scales—have certainly been used widely and are likely to be among the most frequently used measures of depression (H6). However, it is an empirical question to estimate the frequency with which measures of depression are used in *general*.

Second, it is also worthwhile examining the extent to which measures are used disproportionately across treatment and basic science studies. Although the motivation for selecting one measure over another is not always stated explicitly, there is good reason to anticipate differences in the pattern of use between basic science and treatment outcome studies. Treatment studies are frequently conducted for the purpose of regulatory approval or to be comparable to existing treatment studies, whereas in basic science studies, no regulatory concerns exist. Accordingly, it may be hypothesized that (a) a greater number of measures would be used in basic science studies relative to treatment outcome studies (H7), and (b) a greater breadth of measures would have been used in basic science studies than in treatment outcome studies (H8).

Whether a difference in the pattern of use of measures in basic science and treatment studies is important ultimately depends on the quality of those measures. However, studies have shown that the correlation between two of the most frequently used measures of depression, namely the BDI and the HRSD, is modest (Sayer et al., 1993), and that there are a number of concerns with the psychometric properties of the HRSD (Santor & Coyne, 2001b). Should the frequency of use of the BDI and HRSD differ from basic science to treatment outcome studies, then what we know about depressive severity in one domain of research would be overly reliant on one measure that may not perform as optimally as other measures.

Third, examining the frequency of use of different measures also offers important information regarding the extent to which new measures of depression are being adopted by the field. Given the large number of measures and the consistent rate at which new measures are being developed (see Figure 1), it is unclear to what extent new measures are actually being adopted. Although the HRSD is generally regarded as the most frequently used measure of depression, it is nonetheless important to examine the frequency with which the HRSD is being used and the degree to which newer measures that are either superior or at least viable alternatives are actually being implemented by researchers in both basic science and treatment outcome studies. In this study we estimated the frequency of use of measures over a 10-year period of time.

REPRESENTATIVENESS

Finally, we examined the extent to which the most frequently used measures of depression are representative of both (a) measures of depression in general, and (b) the most recently developed measures of depression. In particular, we examined differences in the number and proportion of items devoted to assessing mood, behavioral, somatic, cognitive, and concentration symptoms of depression. Understanding the extent to which the most frequently used measures of depression compare to measures in general provides some indication as to whether the most frequently used measures of depression correspond to our implicit and explicit definitions and operations for measures of depression both (a) generally and (b) most recently. No explicit a priori hypotheses could be formulated.

METHOD

Item Content

To examine the item content of measures of depression (Objective 1), we identified measures of depressive severity from a number of sources, including past reviews of measures, early compendiums, abstract databases (PsychInfo and PubMed), and Internet search engines (www.google.com). We restricted this review to scales and inventories that assessed the severity of depression and excluded diagnostic interviews and unpublished measures.

Although we were able to obtain the original items for the vast majority of scales, some of the earlier measures could not be obtained. In some instances we were able to identify the items from other published articles. However, excluding these scales from our analyses resulted in no appreciable change in the various statistics we computed. We did not include short forms or revisions of the original measure in our review, unless the revision was substantially different. Hybrids of instruments, such as the combined HRSD and BDI proposed by Wechsler, Grosser, & Busfield (1963), were, however, included as separate scales. We also did not include measures that were adapted from adult versions, such as the Center for Epidemiological Studies Depression Scale for Children (Faulstich, Carey, Ruggiero, Enyart, & Gresham, 1986), given that item content developed for one might not be considered independent of the other. Again, we repeated relevant analyses without these adapted scales but found no appreciable differences.

For the purposes of examining the degree of item heterogeneity across different measures of depressive severity, we rated each measure for the number of items assessing distinct symptom domains that appear in Table 1. Although the inter-rater agreement for ratings for the 50 most frequently used measures was high (.95), our review revealed considerable variability in how sad or depressed mood was

TABLE 1
Mean Number of Items Assessing Each Symptom Domain, Given That It Was Assessed

	Number of Items			*Proportion of Total Number of Items*	
Symptom	*M*	*SD*	*Max*	*M*	*SD*
Mood (broad)	1.42	0.97	9	0.21	0.18
Mood (narrow)	1.21	1.13	9	0.17	0.19
Anhedonia	1.16	1.23	7	0.13	0.15
Appetite/weight	0.70	0.84	4	0.06	0.07
Sleep	1.03	1.18	6	0.09	0.09
Irritable	0.70	1.17	10	0.08	0.15
Anxiety	0.82	1.24	9	0.11	0.19
Hopeless	0.93	1.21	7	0.13	0.28
Suicide	0.96	1.14	7	0.10	0.11
Concentration	0.86	1.14	7	0.08	0.10
Energy/fatigue	0.84	1.00	6	0.09	0.10
Worthlessness/guilt	1.84	2.14	14	0.18	0.17
Agitation/retardation	0.78	1.02	6	0.08	0.11
Crying	0.47	0.58	3	0.07	0.11
Withdrawal	0.31	0.70	7	0.04	0.11
No interest in others	0.10	0.31	2	0.01	0.05
Libido	0.18	0.40	2	0.02	0.04
Total items	18.90	15.94	118		

operationalized. Although most measures of depression included items assessing "sad" or "depressed" mood explicitly, many other scales referred only to "feeling down," "feeling blue," or "feeling miserable." Accordingly, we scored whether or not mood was operationalized liberally, including any of the five terms named previously, as well as conservatively, including just "sad" or "depressed."

For our analyses, we analyzed both the number of items used to assess a single symptom domain and the proportion of items allocated to assessing each symptom domain considered. To assess the completeness with which the content domain of depression was assessed, both the frequency with which scales assessed various symptom domains and the number of items used to assess a specific symptom domain were calculated.

Estimating the Frequency of Use

To estimate the frequency with which different measures of depression have been used (Objective 2), we randomly selected from a number of leading journals, published from 1990 to 1999 inclusive (see Appendix). Journals selected included the *American Journal of Psychiatry,* the *Journal of Affective Disorders,* the *British*

Journal of Psychiatry, Archives of General Psychiatry, the *Journal of Nervous and Mental Disease, Psychological Medicine,* the *Journal of Consulting and Clinical Psychology,* the *Canadian Journal of Psychiatry,* and the *Journal of Abnormal Psychology.* Given the relatively smaller number of studies in children and adolescents, we restricted our analyses regarding the frequency of use to studies examining adults only.

To validate the appropriateness of the journals selected, we used EBSCO Host Research Databases to search the PsychInfo database to calculate the total number of articles devoted primarily to the study of depression published in some 89 different journals during the 10 years of interest (1990 to 1999 inclusive). Using the same search method, we calculated the number of articles in each of the 89 journals identified and divided the total for each journal by the total number of articles for all journals, yielding the proportion of articles that each journal devoted to studying depression. Specifically, we used the search term "depression" in all of the basic database fields of the search engine (which is the default setting), namely the title of the journal in question, abstract, subject headings, and keywords. A similar pattern of results was obtained using the PubMed database, as well as using more liberal (e.g., all text fields) or more restricted (e.g., just major subject fields) fields. The 9 journals included in our analyses were ranked in the top 11 journals with the highest proportion of articles on depression. One of the 2 journals not included in our analyses was a specialty journal, *the Journal of Clinical Psychopharmacology.* The only journal not identified on a priori grounds was the *Journal of Clinical Psychiatry.* This confirms that the journals selected were devoted to publishing studies on depression.

From each journal we randomly selected and reviewed 20% of the articles that were identified through PsychInfo. Results of this analysis showed that over 83% of the articles selected for review ($N = 9{,}555$) used a measure of depression. This confirms that the articles selected were devoted to research in depression and were in this sense representative of the journals selected. In summary, we concluded on the basis of these analyses that the journals and articles selected for analysis were representative of research on depression and were as such a good sample from which to estimate how measures of depression have been used.

For each article reviewed, research assistants recorded the various measures of depression that were used, the type of study that was being conducted (basic science study, treatment study, or other), and the population in which the study was conducted. However, of the articles selected for review, not all of the reviewed articles contained measures of depression and not all studies examined a basic science question or an applied topic of interest. Accordingly, for these analyses, we elected to exclude articles that examined only the psychometric properties of depression scales, and we included only those articles that examined depression in a basic science or applied study. Further, we report only results concerning severity measures of depression in adults and excluded interviews and assessment schedules.

RESULTS

Results are presented in three sections: (a) number of measures of depressive severity published in the 80 years up to 2000, (b) item characteristics of measures of depressive severity, and (c) estimated frequency of use.

Number of Measures of Depression

Figure 1 plots the cumulative number of depression scales and inventories published up to the year 2000. Results show a steady rate of growth in the creation of new scales beginning in the 1960s, which only in the past 5 years has shown any sign of slowing. Most scales have been published in the past 20 years. This figure does not include any of the numerous short forms or scale revisions, nor diagnostic interviews, which were outside the scope of the current study.

Item Characteristics

Frequency with which symptom domains were assessed. Figure 2 shows the percentage of scales assessing each of a wide range of symptom domains that are typically assessed in measures of depression. A repeated measures analysis of variance showed that the proportion with which symptom domains were assessed was not equitable across symptom domains, $F(15, 4020) = 73.25$, $p < .0001$. Mood, anhedonia, and worthlessness were among the most frequently measured symptom domains. Subsequent analyses comparing the frequency with which depressed mood, anhedonia, and worthlessness was assessed showed that worthlessness was not assessed less frequently than was anhedonia, $F(1, 268) = 3.14$, $p < .08$. Of interest, mood was not uniformly operationalized as feeling "sad" or "depressed." Approximately 20% of measures of depression operationalized mood with terms other than feeling "sad" or "depressed."

Number of items used to assess symptom domains. In addition to assessing a different number of symptom domains, results showed that different symptom domains were assessed in different ways. Both the mean number of items used to assess an individual symptom and the proportion of items devoted to assessing an individual symptom are presented in Table 1. Results showed that (a) the mean number of items used to assess an individual symptom, $F(15, 4020) = 51.00$, $p < .0001$, and (b) the proportion of items assessing an individual symptom, $F(15, 4020) = 42.00$, $p < .0001$, differed significantly across the 16 symptom domains assessed in the analysis (H1).

Supplementary analyses showed that the mean *number* of items used to assess worthlessness ($M = 1.84$, $SD = 2.14$) was greater, $F(1, 268) = 50.08$, $p < .0001$, than the mean number of items used to assess symptom domains on average ($M = 1.06$,

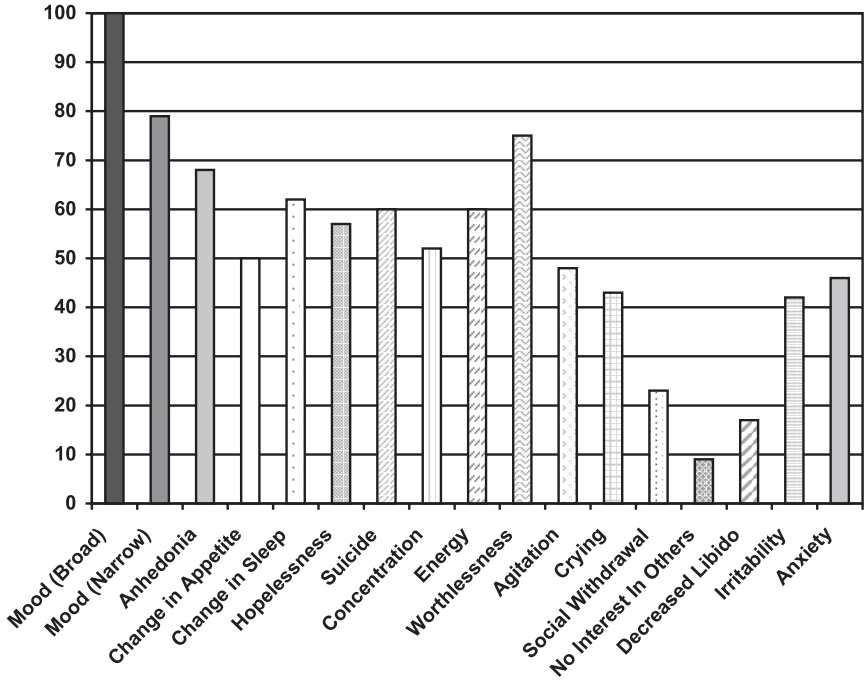

FIGURE 2 Percentage of measures assessing a variety of symptom domains. Sad or depressed mood and worthlessness are most frequently assessed. Reduced libido is the least often assessed. As many as 20% of measures do not assess sad or depressed mood explicitly. Results for decreased libido were computed only for adolescent and adult measures of depression.

SD = 0.68), and that the mean *proportion* of items used to assess worthlessness (*M* = 0.18, *SD* = 0.17) was greater, $F(1, 268) = 40.31$, $p < .0001$, than the mean proportion of items used to assess a symptom domain in general, which was 0.11 (H4). Supplementary analyses also showed that the mean number of items used to assess worthlessness ($M = 1.84$, $SD = 2.14$) was greater, $F(1, 268) = 21.87$, $p < .0001$, than the mean number of items used to assess anhedonia ($M = 1.16$, $SD = 1.12$), and that the mean proportion of items used to assess worthlessness ($M = 0.18$, $SD = 0.17$) was greater, $F(1, 268) = 12.12$, $p < .0001$, than mean the proportion of items used to assess anhedonia ($M = 0.13$, $SD = 0.15$).

These findings show that despite being intended as measures of the same underlying construct, the degree of variability across measures, in terms of the number of items used to assess different symptoms, is substantial. Results suggest that the symptoms primarily assessed by measures of depression are worthlessness, depressed mood, and anhedonia, even though only depressed mood and anhedonia are considered core symptoms of depression.

Differences in adult and adolescent measures of depression. We hypothesized that adult and adolescent measures of depression would operationalize irritability differently, either by (a) assessing irritability more frequently, (b) with a greater number of items in total, or (c) as a proportion of the total number of items in the scale. Analyses revealed no differences in (a) the proportion of adult and adolescent scales assessing irritability, (b) the number of items used to assess irritability computed either as the total number or items assessing irritability, or (c) as a proportion of total number of items in the scale (H2).

Changes over time. We also examined the extent to which item and scale characteristics varied over time. Measures were grouped into 3 time periods, 1918 to 1969 ($n = 50$), 1970 to 1989 ($n = 120$), and 1990 to 2000 ($n = 117$), and analyzed in a general linear model with two class variables coding for time. Dependent variables were (a) total number of items assessing a symptom domain, (b) the proportion of items assessing a symptom domain, and (c) whether or not a symptom domain was at all assessed. Given the large number of statistical tests (16 symptom domains × 3 measures), results showed remarkable consistency in how depression has been operationalized over the past 80 years. However, some significant differences were observed, which are summarized in Table 2. Of the differences that were significant, consistent increases over time were observed for the proportion of items assessing suicidal thought and behavior, for the number and proportion of items assessing loss of interest, and for the number of scales assessing loss of inter-

TABLE 2
Changes in Item Characteristic Over Time

	Time Period							
	Before 1970		*1970–1985*		*1985–2000*			
	M	*SD*	*M*	*SD*	*M*	*SD*	*F*	*p*
Number of items								
Loss of interest	0.94	1.27	1.33	1.27	1.04	1.11	3.13	<.04
Anxiety	1.24_{a}	1.67	0.75_{b}	1.08	0.67_{b}	1.11	4.08	<.02
Proportion of items								
Agitation	0.11_{a}	0.11	0.09_{ab}	0.12	0.06_{b}	0.88	4.81	<.01
Worthlessness	0.18	0.15	0.15	0.15	0.20	0.17	3.86	<.02
Suicide	0.08_{a}	0.08	0.08_{a}	0.10	0.12_{b}	0.12	4.68	<.01
Irritability	0.11_{a}	0.18	0.06_{b}	0.09	0.09_{ab}	0.16	3.19	<.04
Loss of interest	0.09_{a}	0.11	0.15_{b}	0.11	0.12_{ab}	0.14	3.63	<.03
Symptom domain								
Worthlessness	0.80	0.40	0.67	0.48	0.80	0.40	3.44	<.04
Loss of interest	0.48_{a}	0.11	0.78_{b}	0.41	0.65_{b}	0.48	8.08	<.01

Note. Means with different subscripts differed significantly at $p < .05$ on post hoc tests.

est at all. Consistent decreases were observed only for the number of items used to assess anxiety and the proportion of items assessing agitation (H3).

Frequency of Use

Results of our analyses concerning the frequency with which measures of depressive severity were actually used for both basic science and treatment outcome studies are presented in Figure 3. Results in Figure 3 show that the BDI and the HRSD are among the most frequently used measures of depressive severity in both basic science and treatment outcome studies. Together the BDI and the HRSD were used in approximately 42% of basic science studies and in approximately 63% of all treatment outcome studies (H6). Analyses showed that the distribution of measures across basic science and treatment outcome studies differed significantly overall, $\chi^2(5) = 30.31$, $p < .0001$ (H7), and that this overall effect could be attributed to significant differences between basic science and treatment outcome studies for the BDI, $\chi^2 = 5.23$, $p < .02$; CES-D, $\chi^2 = 7.29$, $p < .007$; HRSD, $\chi^2 = 12.20$, $p < .0005$; and other measures, $\chi^2 = 5.07$, $p < .02$. However, no significant difference for the Montgomery Asberg Depression Scale was found.

Results also showed that a greater variety of depressive severity measures were used in basic science studies ($n = 97$) than in treatment outcome studies ($n = 33$; H8). However, only six measures of depression severity were used in more than 40% of studies reviewed.

Representativeness of the Most Frequently Used Measures

To assess whether the most frequently used measures are representative of the manner in which depressive symptoms have been operationalized in general, we compared the manner in which the BDI, CES-D, Montgomery Asberg, and HRSD

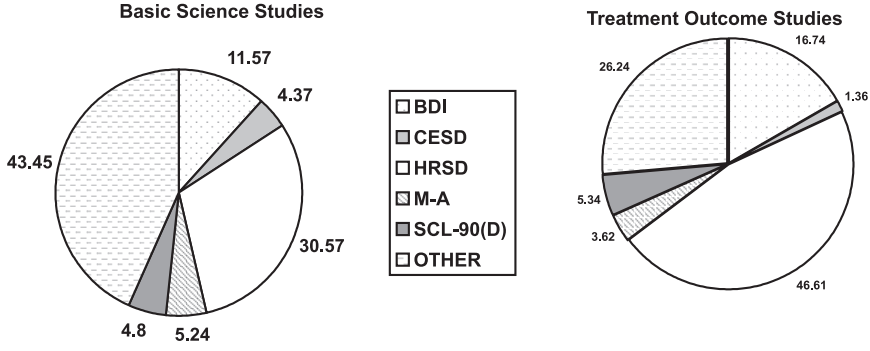

FIGURE 3 Pie graphs showing the proportionate frequency of use of measures of depression. In total some 70 different measures of depression were used in the past 10 years. However, only 6 measures were used with any frequency.

differed from depression scales in general, in terms of the proportion of mood symptoms (depressed mood and irritability), behavioral symptoms (suicide and anhedonia), somatic symptoms (appetite disturbance, sleep disturbance, low energy, and psychomotor retardation or agitation), cognitive symptoms (hopelessness and worthlessness), and concentration symptoms (poor concentration and decision making).

Differences were assessed in a series of *t* tests comparing the average score for all scales with the proportion of items used to assess each domain of interest for the BDI, CES-D, HRSD, M-A, and SCL90-D. Results are presented in Table 3 and show the extent to which each of the selected measures differs from the entire population of published measures. Results show that the BDI is generally more representative of depression scales in general, differing only slightly in the proportion of items used to assess mood and cognitive symptoms. Indeed, the BDI devotes more items to assessing features of worthlessness and hopelessness and fewer items to assessing mood than other measures in general. In contrast, results showed that the HRSD devotes more items to assessing somatic symptoms than measures in general and fewer items to assessing cognitive features than measures in general.

We also examined differences between these measures and measures created in only the past 15 years. Results showed that a number of differences found with respect to measures in general no longer existed. Differences found with respect to the entire group of measures *not* found in measures created in the past 15 years are highlighted in italics. These results suggest that the BDI–II is even more representative of measures created in the past 15 years. Results for the HRSD remained unchanged.

Summary

Results showed (a) that the mean number and proportion of items used to assess different symptoms differed significantly across symptom domain (see H1), (b)

TABLE 3
Representativeness of Measures

	Score						
	M	*SD*	*BDI*	*CES-D*	*HRSD*	*M-A*	*SCL90*
Proportion of Items Assessing							
Mood symptoms	0.19	0.31	–0.09**	–0.08**	–0.12**	*ns*	–0.11**
Behavioral symptoms	0.16	0.22	*ns*	–0.06**	–0.04**	0.14**	0.07**
Somatic symptoms	0.24	0.40	*ns*	–0.14**	0.17**	*ns*	–0.08*
Cognitive symptoms	0.20	0.32	*0.09***	–0.10**	–0.14**	–0.10**	*ns*
Concentration symptoms	0.05	0.12	*ns*	*–0.01***	–0.05**	*0.05***	–0.05**

*$p < .05$. **$p < .01$.

that the total number and proportion of items assessing irritability did not differ between adult and adolescent scales (see H2), (c) that only a few changes in the characteristics of measures emerged over time (see H3), (d) that the mean number and proportion of items used to assess worthlessness was greater than the mean number and proportion of items used to assess a symptom in general (see H4) but did not increase consistently over time (see H5), (e) that the BDI and HRSD are, in fact, among the most widely used measures of depression (see H6), and (f) that a greater number (see H7) and breadth (see H8) of measures are used in basic science studies than in treatment outcome studies. Results also showed that the BDI was more representative of depression measures in general.

DISCUSSION

The purpose of the study was to examine the item characteristics of the various scales that have been developed and published over the past several decades and to estimate the frequency with which different measures of depression have been used in both basic science and treatment outcome studies.

Item Characteristics

Results also showed that the manner in which symptoms have been operationalized in general varies considerably and that this variability is characteristic of measures of depression in general. Consistent with the earlier review by Snaith (1993), which examined a small number of frequently used measures, results in this study showed that both the number of items allocated to assessing different symptoms and whether a symptom domain was assessed at all varied significantly across measures of depression in general.

As anticipated, core symptoms of depression, namely depressed mood and anhedonia, were assessed most frequently on measures of depressive severity (Figure 2). However, results also showed that the mean number of items and the mean proportion of items used to assess worthlessness were greater than for anhedonia. This set of findings suggests that worthlessness is operationalized in a manner consistent with other core symptoms of depression, even though it is not generally considered a core symptom of depression for diagnostic purposes. Historically, feelings of worthlessness have been considered core features of depression by individuals such as Bibring (1953) and Beck (1963). These data raise questions concerning the manner in which depression is conceptualized. For some, the core features of depression may be restricted to mood and anhedonia. However, our data show that most scales have been developed in a manner consistent with viewing worthlessness as a core feature of depressive severity. Arguably, the manner in which depression is being implicitly defined and

operationalized may warrant affording symptoms such as worthlessness more prominence or status than they currently hold.

Changes Over Time

We also examined changes in the item characteristics of scales over time. In general, scales showed remarkable consistency over time. Differences that were significant were generally few. Indeed, many of the differences observed were fluctuations that may reflect the manner in which the time periods were chosen rather than real changes in how depressive severity is being operationalized. Only one hypothesized difference was confirmed, namely that the number of items used to assess anxiety has generally decreased, which is consistent with changes in formal definitions of depression as characterized in different versions of the *DSM*. One unexpected difference was found. Results showed (a) that the number of items used to assess loss of interest has increased, (b) that the proportion of items used to assess loss of interest has also increased, and (c) that the proportion of scales assessing loss of interest at all has increased over the past 80 years.

Although *DSM* criteria have changed over the past several decades, measures assessing depression-related constructs have not followed *DSM* criteria, which may account for the failure to find corresponding changes in depressive severity measures. For Bridgeman (1927), this change in operational criteria would represent a change in how depression is defined and conceptualized. Indeed, substantial changes in how depression is assessed stand to threaten the continuity of knowledge. If changes in how a construct is operationalized change the psychometric properties of the scale, such as the mean score, inferences about how depressed a person is, as well as the efficacy of treatments, would also be affected.

Number of Scales

Despite the large number of scales constructed in the past 80 years, relatively few are used currently and even fewer are used with any notable frequency. Although some 280 measures of depressive severity have been published in the past 80 years, only some 70 scales have been used in the recent past, and the majority of studies with adults have relied on just 1 of 5 different scales. It is notable that the most frequently used scales, the BDI, CES-D, HRSD, M-A, and SCL90-D, were developed more than 20 years ago. This underscores the difficulty of introducing a new measure of depression and seeing it adopted by the field. Despite the problems that have been identified with the HRSD and the large number of scales that have been published since its introduction, some of which were developed specifically to address the problems with the HRSD, the

HRSD remains one of the most frequently used measures of depressive severity, especially in applied research. Adopting a new measure of depression will require the commitment of a large number of stakeholders, including researchers, clinicians, and industry, as well as regulatory bodies. Without broad-scale support for the development and implementation of a new measure, it is unlikely that any new measure will be adopted widely.

Frequency of Use

Results also showed a difference in the frequency with which these measures were used in both basic science and treatment outcome studies. Measures of depression such as the BDI, HRSD, M-A, and SCL90-D were used less frequently in the sample of basic science studies than in the sample of treatment outcome studies; more measures of depression were used in basic science studies than in treatment outcome studies. Ideally, the various relations that exist among depressive severity and other related constructs should be substantiated across a number of measures of depressive severity, rather than be dependent on just a few measures of depressive severity. If findings with one measure do not generalize to other valid measures of depressive severity, the robustness of the effect may be questionable and likely idiosyncratic to the measure.

Our findings also showed that a greater number of measures are being used to examine the correlates and effects of depressive severity in basic science studies (n = 97) than in treatment outcome studies (n = 33). Given that most treatment outcome studies have been conducted with either the HRSD or the BDI (or both), it is fair to argue that much of what we know about the effectiveness of treatments depends on quality of just the BDI and HRSD. In contrast, results in Figure 3 imply that much of what we know about depressive severity through basic science studies does not depend nearly as strongly on the quality of the BDI and HRSD.

Whether differences in (a) the frequency with which the core measures are used across basic science and treatment outcome studies, or (b) the number of measures used in general are meaningful depends ultimately on the quality of those measures. Indeed, one might argue that use of a large number of measures might in fact complicate drawing meaningful conclusions. Given that there is now good evidence to show that the principal measures used to evaluate treatment outcomes, namely the BDI and HRSD, correlate only modestly (about .52; Sayer et al., 1993), it is at least worthwhile to carefully consider the quality and appropriateness of the most frequently used measures of depression.

Representativeness

Last, we examined the extent to which the most frequently used measures were representative of measures in general. Results showed that the BDI was generally

representative of measures of depression, particularly of those measures developed and published in the last 15 years. In contrast, results showed that the HRSD and the CES-D differed from measures in general in all of the domains assessed. The HRSD allocates more items to assessing somatic symptoms and fewer items to all other symptom groups. In contrast, the CES-D includes a number of items that are unlikely to be specific to depressive disorders, such as the perceptions of others (Item 15 and Item 19), talkativeness (Item 13), or comparisons with others (Item 4). These results suggest that two of the primary measures, namely the HRSD and the CES-D, on which much of what we know about basic science and treatment outcome studies depends, are not representative of how measures of depression have been operationalized.

Limitations

Although we selected journals that publish large numbers of articles on depression, the potential for biased results arising from the manner in which articles were selected must be acknowledged. To our knowledge, this is the first analysis of its kind with measures of depression, and in this regard many of the potential biases may not be known or well understood. Had we included other journals, then we may have had different results. However, the journals from which articles were drawn were selected because they have published large numbers of papers on depression. Specialty journals such as *Psychological Assessment* may publish a different type of study or depression measure, but the impact of specialty journals on the overall findings would likely be small. Still, the possibility for bias is real. Clearly, further analysis and replication is recommended. Finally, our analyses focused exclusively on measures of depressive severity, the results of which may not necessarily generalize to clinician-rated measures of depression.

Conclusion

The standardized assessment of depressive severity remains an important goal of measurement. Although there are good reasons for constructing different measures of depression, the large number of often diverse measures created over the past 80 years creates its own challenge to one of the main goals of measurement, namely to standardize assessment and streamline decision making. Continued progress in the assessment of depression depends on understanding the manner in which depression is conceptualized and operationalized, the manner in which scales purporting to assess the same underlying construct actually differ, and the impact that such differences have on performance (cf. Santor 2005; Santor & Ramsay, 1998). Despite the continuous development and validation of new measures of depression, the field continues to rely on measures created some 25 to 30 years ago, some of which have considerable shortcomings and are not necessarily

representative of how depressive severity is typically assessed. This illustrates the difficulty of seeing new measures introduced and adopted, particularly in treatment outcome studies. Indeed, without broad-scale support for the development and implementation of a new measure by the important stakeholders, including industry and regulatory bodies, it is unlikely that any new measure will be adopted widely.

ACKNOWLEDGMENTS

We are grateful for the assistance of Melissa Burgess in preparing the manuscript. Copies of the complete reference list of measures identified and measures database are available on request.

REFERENCES

American Psychiatric Association. (1952). *Diagnostic and statistical manual of mental disorders*. Washington, DC: Author.

American Psychiatric Association. (1968). *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, DC: Author.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., revised). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Arieti, S., & Bemporad, J. R. (1980). The psychological organization of depression. *American Journal of Psychiatry, 137,* 1360–1365.

Beck, A. T. (1963). Thinking and depression: Idiosyncratic content and cognitive distortions. *Archives of General Psychiatry, 9,* 324–333.

Beck, A. T., Steer, R. A., &. Brown, G. K. (1996). *Manual for the Beck Depression Inventory II*. San Antonio, TX: Psychological Corporation.

Beck, A. T., Steer, R. A., & Garbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review, 8,* 77–100.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561–571.

Bibring, E. (1953). The mechanism of depression. In P. Greenacre (Ed.), *Affective disorders* (pp. 13–48). New York: International Universities Press.

Blatt, S. J. (1974). Levels of object representation in anaclitic and introjective depression. *Psychoanlytic Study of the Child, 29,* 107–157.

Bowlby, J. (1969). *Attachment and loss* (Vol. 1). Harmondsworth, England: Penguin.

Bridgeman, P. W. (1927). *The logic of modern physics*. New York: Macmillan.

Carnap, R. (1936). Testability and meaning. *Philosophy of Science, 3,* 420–468.

Carnap, R. (1937). *The logical syntax of language*. New York: Harcourt Brace.

Coyne, J. C. (1976). Toward an interactional description of depression. *Psychiatry, 39,* 28–40.

Coyne, J. C. (1994). Self-reported distress: Analog or ersatz depression? *Psychological Bulletin, 116,* 29–45.

Faulstich, M. E., Carey, M. P., Ruggiero, L., Enyart, P., & Gresham, F. (1986). Assessment of depression in childhood and adolescence: An evaluation of the Center for Epidemiological Studies Depression Scale for Children (CES-DC). *American Journal of Psychiatry, 143,* 1024–1027.

Flett, G. L., Vredenburg, K., & Krames, L. (1997). The continuity of depression in clinical and nonclinical samples. *Psychological Bulletin, 121,* 395–416.

Freud, S. (1917). Mourning and melancholia. In J. Strachey (Ed. and Trans.), *The standard edition of the complete psychological works of Sigmund Freud* (Vol. 14, pp. 243–258). London: Hogarth Press.

Garret, H. E., & Schenk, M. R. (1928). A study of the discriminative value of the Woodworth Personal Data Sheet. *Journal of General Psychology, 1,* 459–471.

Gilbert, P. (1992). *Depression: The evolution of powerlessness*. New York: Guilford.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry, 23,* 56–62.

Kendall, P. C., & Flannery-Schroeder, E. C. (1995). Rigor, but not rigor mortis in depression research. *Journal of Personality and Social Psychology, 68,* 892–894.

Kendall, P. C., Hollon, S. D., Beck, A. T., & Ingram, R. E. (1987). Issues and recommendations regarding the use of the Beck Depression Inventory. *Cognitive Therapy and Research, 11,* 289–299.

Lewinsohn, P. M., Youngren, M. A., & Grosscup, S. J. (1979). Reinforcement and depression. In R. A. Depue (Ed.), *The psychobiology of depressive disorders: Implications for the effects of stress* (pp. 291–316). New York: Academic.

Nezu, A. M., McClure, K. S., Meadows, E. A., & Ronan, G. F. (2000). *Practitioner's guide to empirically based measures of depression.* New York: Kluwer Academic.

Price, J. C., Sloman, L., Gardner, R. Jr., Gilbert, P., & Rohde, P. (1994). The social competition hypothesis of depression. *British Journal of Psychiatry, 164,* 309–315.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385–401.

Santor, D. A. (2005). Using and evaluating psychometric measures: Practical and theoretical considerations. In J. Miles & P. Gilbert (Eds), *A handbook of research methods for clinical and health psychology* (pp. 95–109). Oxford, England: Oxford University Press.

Santor, D. A., & Coyne, J. C. (2001a). Evaluating the continuity of symptomatology between depressed and nondepressed individuals. *Journal of Abnormal Psychology, 110,* 216–225.

Santor, D. A., & Coyne, J. C. (2001b). Examining symptom expression as a function of symptom severity: Item performance on the Hamilton Rating Scale for Depression. *Psychological Assessment, 13,* 127–139.

Santor, D. A., & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment, 10,* 345–359.

Sayer, N. A., Sackhein, H. A., Moeller, J. R., Prudic, J., Devanand, D. P., Coleman, E. A., et al. (1993). The relations between observer-rating and self-report of depressive symptomatology. *Psychological Assessment, 5,* 350–360.

Snaith, P. (1993) What do depression rating scales measure? *British Journal of Psychiatry, 163,* 293–298.

Suppes, F. (1977). The search for philosophic understanding of scientific theories. In F. Suppes (Ed.), *The structure of scientific theories* (2nd ed., pp. 3–257). Chicago: University of Illinois Press.

Tennen, H., Hall, J. A., & Affleck, G. (1995a). Depression research methodologies in the Journal of Personality and Social Psychology: A review and critique. *Journal of Personality and Social Psychology, 68,* 870–884.

Tennen, H., Hall, J. A., & Affleck, G. (1995b). Rigor, rigor mortis, and conspiratorial views of depression research. *Journal of Personality and Social Psychology, 68,* 895–900.
Vredenburg, K., Flett, G. L., & Krames, L. (1993). Analog versus clinical depression: A clinical reappraisal. *Psychological Bulletin, 113,* 327–344.
Wechsler, H., Grosser, G. H., & Busfield, B. L., Jr. (1963). The depression rating scale. *Archives of General Psychiatry, 9,* 46–55.

APPENDIX
Journal Selection Pool

Addiction
American Journal of Geriatric Psychiatry
American Journal of Psychiatry
American Psychologist
Archives of General Psychiatry
Assessment
Behavioral and Cognitive Psychotherapy
Behavioral Neuroscience
British Journal of Clinical Psychology
British Journal of Psychiatry
British Journal of Psychotherapy
British Journal of Social and Clinical Psychology
Canadian Journal of Psychiatry
Chinese Journal of Clinical Psychology
Contemporary Psychology
Depression
Depression and Anxiety
Depression and Stress
Developmental Psychology
Emotion
Experimental and Clinical Psychopharmacology
Health Psychology
International Journal of Geriatric Psychiatry
International Journal of Psychiatry in Clinical Practice
International Journal of Psychiatry in Medicine
International Journal of Social Psychiatry
International Journal of Stress Management
International Review of Psychiatry
Journal of Personality
Journal of Abnormal Child Psychology
Journal of Abnormal Psychology
Journal of Affective Disorders
Journal of Anxiety Disorders
Journal of Applied Psychology
Journal of Behavioral Medicine
Journal of Child and Adolescent Psychopharmacology
Journal of Child Psychology and Psychiatry and Allied Disciplines
Journal of Child Psychotherapy
Journal of Clinical Child Psychology
Journal of Clinical Psychoanalysis
Journal of Clinical Psychology
Journal of Clinical Psychology in Medical Settings
Journal of Clinical Psychopharmacology
Journal of Consulting and Clinical Psychology
Journal of Contemporary Psychotherapy
Journal of Counseling Psychology
Journal of Family Psychology
Journal of Geriatric Psychiatry and Neurology
Journal of Nervous and Mental Disease
Journal of Neurology, Neurosurgery, and Psychiatry
Journal of Neuropsychiatry & Clinical Neurosciences
Journal of Personality and Social Psychology
Journal of Personality Assessment
Journal of Personality Disorders
Journal of Psychiatric Research
Journal of Psychiatry and Neuroscience
Journal of Research in Personality
Journal of Social and Clinical Psychology
Journal of the American Medical Association
Journal of the Psychiatric Association of Thailand
Journal of Traumatic Stress
Journal of Youth and Adolescence

New England Journal of Medicine
Prevention and Treatment
Primary Care Psychiatry
Professional Psychology: Research and Practice
Psychiatric Annals
Psychiatric Bulletin
Psychiatry and Clinical Neurosciences
Psychological Assessment
Psychological Bulletin
Psychological Medicine
Psychological Methods
Psychological Review
Psychology and Aging
Psychopharmacology
Psychosomatic Medicine
Psychotherapy: Theory, Research, Practice, Training
Suicide and Life-Threatening Behavior
The Journal of Clinical Psychiatry